Séminaire Annuel du CERIDES / 08 et 09 juillet 2021

Du crawling intelligent au fact-checking participatif: le cas du projet de recherche Check4Decision

Dr Edouard Ngor SARR

Maitre Assistant – CAMES

Université Assane SECK de Ziguinchor-UASZ

Check4Decision Project, ML4FC, CERIDES UCAO

Plan

- **Définitions**
  - Web Scraping & Crawling Intelligent
  - Fact-checking
- **Contexte**
- **Problématique**
- **Complexité de la question**
  - Différentes phases du processus et les travaux realisés
- **Etat de l'art**
  - Projets de FC automatisés
- **Projet de recherche check4Decision**
  - Axes, Equipe de recherche, Architecture du Système
  - Résultats
  - Travaux en cours

Définitions

Web Scraping

- **Grattage de données web**
- Collecte automatique des données web en connaissant à priori la structure des Pages cibles
- Nécessité de connaitre la structure de la page web concernée

Fact-checking participatif

- Implication des internautes (le lecteur) dans la vérification de l'information

Web Crawling

- **Exploration de données web**
- Parcourir un ou un ensemble d'URL (Index)
- Atteindre la profondeur maximale d'une page
- Explorer de manière itérative sans connaissance préalable de la structure des pages cibles
- Collecte des données qui satisfaits la condition

Intelligent Web Crawling

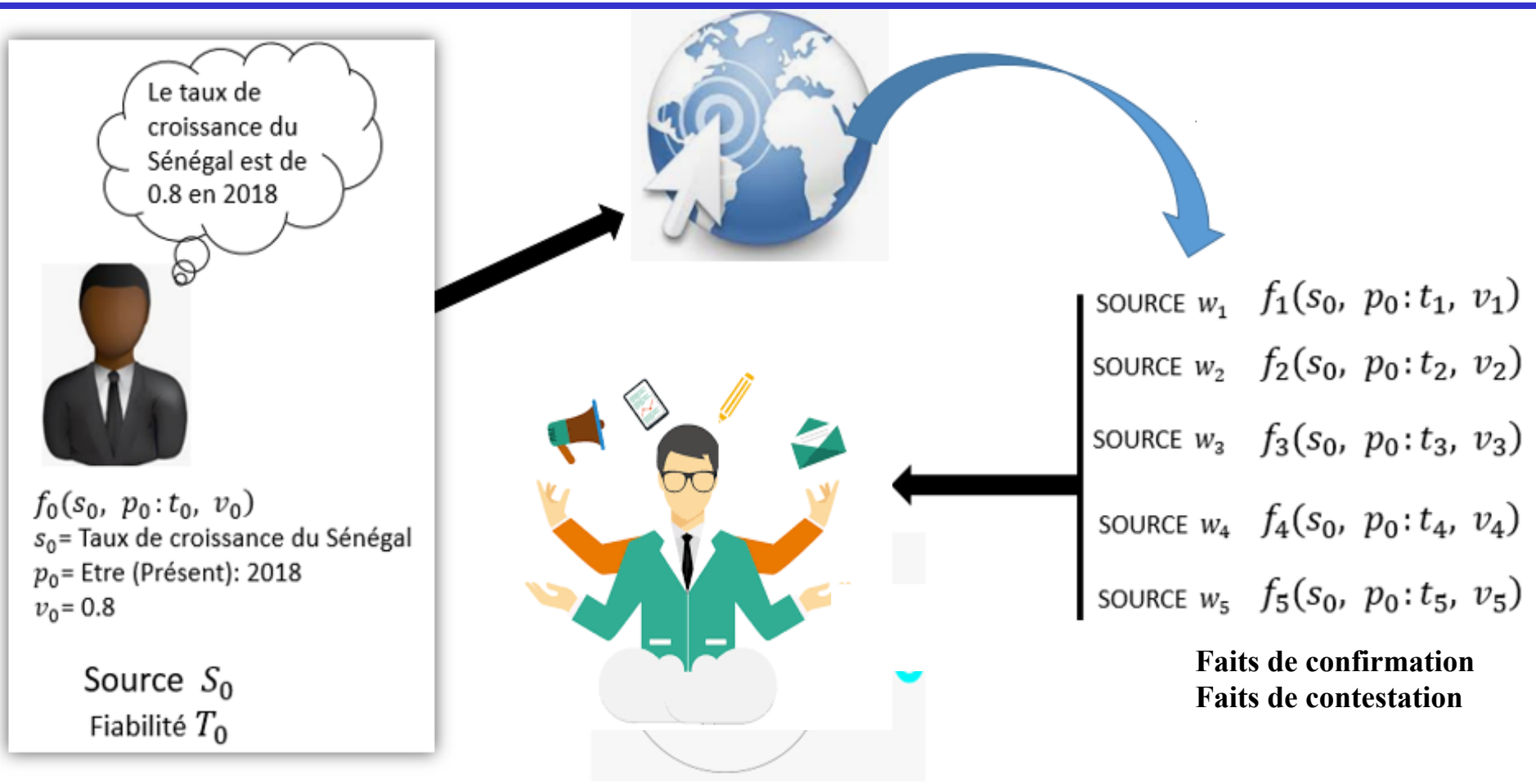
- Utilisation des techniques du machine Learning pour effectuer un crawling web

Fact-checking

- Identifier la vérité, parmi un ensemble de propositions (faits) potentiellement contradictoires.
  - Pré Fact-checking: Apparu au USA vers 1923 ave le Magazine TIME
  - Post Fact-checking: entre 1960 et 1970 avec «watchdog journalism » ou journalisme de surveillance (Affaire du Watergate).

Contexte

Fact-checking Manuel



Fastidieux, Lent, Non fiable

Problématique

Automatisation du Fact-checking ?

Collecte & Agrégation de données web Tagging (TALN)

OUTILS & LOGICIELS

SOURCE w_1 $f_1(s_0,\ p_0 : t_1,\ v_1)$

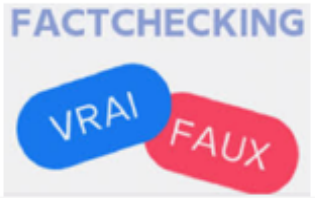
SOURCE w_2 $f_2(s_0,\ p_0 : t_2,\ v_2)$

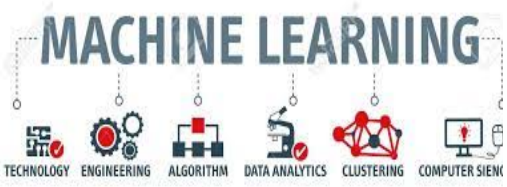
SOURCE w_3 $f_3(s_0,\ p_0 : t_3,\ v_3)$

SOURCE w_4 $f_4(s_0,\ p_0 : t_4,\ v_4)$

SOURCE w_5 $f_5(s_0,\ p_0 : t_5,\ v_5)$

Faits de confirmation
Faits de contestation





Complexité de la question
C1: Acquisition des données

- **C11: Modélisation des articles, faits et commentaires**

News Articles and Facts Modeling: Linked Data or Linked Fact
EN **Sarr**, SN Mbaye, L Faty, O Sall... - ... on Advances in ..., 2020 - ieeexplore.ieee.org
A press article is a collection of statements (opinions and/or allegations) that are a priori non-factual (the facts), ie an opinion on a specific subject. This article seeks to address the challenging issue of modeling news articles and facts. The authors also discuss the ...
☆ Autres articles

News Comments Modeling for Opinion Mining: The Case of Senegalese Online Press
L Faty, M Ndiaye, EN **Sarr**, O Sall... - ... on Advances in ..., 2020 - ieeexplore.ieee.org
Opinion Mining is the processing textual contents from on-line discussions in order to highlight the opinions of net surfers. In the process, the identification of the comments through a model and a formalism which encompass the methods and the data is critical step ...
☆ Cité 1 fois Autres articles

- **C12: Collecte et agrégation des données face à l'Hétérogénéité des sources, des formats, des structures**
  - Web scraping et le Crawling Intelligent

Web scraping: state-of-the-art and areas of application
R Diouf, EN **Sarr**, O Sall, B Birregah... - ... Conference on Big ..., 2019 - ieeexplore.ieee.org
Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, database or CSV file. However, in addition to be a very complicated task, Web Scraping is resource and time consuming ...
☆ Cité 20 fois Autres articles Les 3 versions

FactExtract: automatic collection and aggregation of articles and journalistic factual claims from online newspaper
EN **Sarr**, S Ousmane, A Diallo - 2018 Fifth International ..., 2018 - ieeexplore.ieee.org
Nowadays the web is very rich that's why it is the principal corpus of fact checkers. These journalists in their quest for the truth daily scan manually a lot of sources in search of relevant informations. But in the data-journalism context where data are always increasing in ...
☆ Cité 11 fois Autres articles

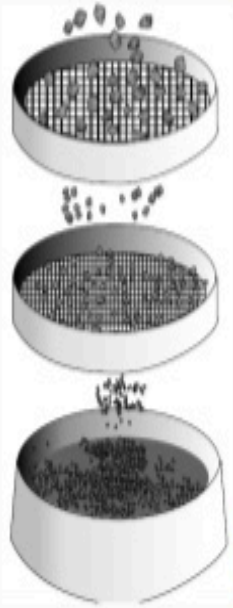
Domaines
- **Marketing**
  - Veille concurrentielle
- **Tourisme & Transport**
  - Comparateurs de prix
- **Sécurité**
  - Surveillance Utilisateurs

Complexité de la question
C2: Identification des faits factuels

Pour l'optimisation des processus = Disposer de méthodes intelligentes pour juger ou non de la pertinence à la vérification

Ground Truth Collection

iDiR



20788 sentences
374 coders
76552 labels

- 20 months, 374 coders, ~$4,000 paid
- 30 training sentences
- 1032 screening sentences (731 NFS, 63 UFS, 238 CFS) to detect spammers & low-quality coders

86 top-quality coders
52333 labels

Majority voting
20617 admitted sentences

Class	Count	Percentage
CFS	4849	23.52%
UFS	2097	10.17%
NFS	13671	66.31%

http://www.engineersdaily.com/2014/03/basics-of-soil-compaction.html

18

Les NFS (Non- Factual Sentence) : affirmations non factuelles

UFS (Unimportant Factual Sentence): phrases factuelles mais non pertinentes à la vérification

Les CFS (Check-worthy Factual Sentens) : phrases factuelles et dont l'étude de la véracité est trés importante

Complexité de la question
C3 Étiquetage des faits

Un fait est une affirmation factuelle qui a déjà été vérifiée ou qui peut l'être sur la base de sources fiables.

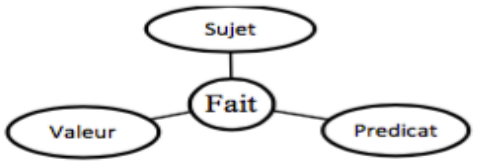


Figure 3. Représentation visuelle du fait

Automatic Segmentation and tagging of facts in French for automated fact-checking

EN **Sarr**, O Sall, A Maiga, L Faty... - 2018 IEEE International ..., 2018 - ieeexplore.ieee.org

In recent years, automatic natural language processing (NLP) has made considerable progress in terms of performance. Nevertheless, to undertake a linguistic analysis of the facts in French remains a real problem today. On the one hand, current taggers do not match ...

☆ 99 Cité 4 fois Autres articles

Part-Of-Speech Tagging in French: State-of-the-Art and Obstacles

EN **Sarr**, O Sall, L Faty - 2020 Seventh International Conference ..., 2020 - ieeexplore.ieee.org

POS Tagging (Part-Of-Speech Tagging) is considered as a basic step for automatic natural language processing. It consists in identifying the grammatical category of each word in a text. It is tricky for a machine to understand the grammatical form of words in some ...

☆ 99 Autres articles

Complexité de la question
C4: Estimation de la véracité & la fiabilité

- **La fiabilité d'une source exprime la probabilité de la source à fournir des faits vrais.**
- **La véracité** d'un fait exprime la probabilité pour que le fait soit vrai.
  - Calculée en utilisant la fiabilité des sources ayant exprimées le fait.
  - Modélisée sous la forme d'un score compris entre 0 et 1

MajorityVoting
TruthFinder
Cosine
2-Estimates
3-Estimates
SimpleLCA
GuessLCA
AccuSim
Depen
Accu
AccuNoDep
LTM
MLE
mean
stddev
min
max

SenFact Algorithm: Fact-checking by the confrontation of opinions

EN **Sarr**, O Sall, A Diagne - 2017 13th International Conference …, 2017 - ieeexplore.ieee.org

The pluralism of media and the unbridled running in production have resulted in a considerable increase of the volume of treated and produced data, but also the proliferation of wrong information. The internet-user is so confronted with two major problems: the …

☆ 99 Cité 5 fois Autres articles

SnVera: A New Algorithm for Automation of Fact-Checking in Web Journalism Context

EN **Sarr**, O Sall, A Diallo - 2018 Fifth International Conference …, 2018 - ieeexplore.ieee.org

In this paper, we present a fact-checking algorithm named SnVera based on the confrontation of opinions. We mean by confronting opinions, using the identical point of view and controversial point of view of the sources of the corpus, as essential elements in the …

☆ 99 Cité 2 fois Autres articles

Etat de l'art

Bref aperçu des projets de FC automatisé

Youtube DataViewer

In video veritas

Google Fact Check Tools

- Journal *Le Monde* en France (https://www.lemonde.fr/verification/)

- **Du NFS aux Etats Unis**
- Un modèle d'IA pour détecter automatiqueme nt les affirmation à vérifier.
- https://idir.uta.e du/claimbuster/

InVid et CrossCheck lancés par le réseau de fact-checking First Draft News (https://fr.firstdraftnew s.com/) et financés par Google News Lab. http://invid.condat.de/

	Mode d'utilisation	Type de données en entrée				
	Extension navigateur	Plateforme en ligne	Image	Audio/Vidéo	Email	Texte
VERA		✓				✓
FactMinder		✓				✓
Claimbuster		✓				✓
Truth Teller		✓		✓		✓
MapChecking		✓	✓			
CrossCheck		✓				✓
Hoaxbuster		✓			✓	
Decodex	✓					✓
Check News		✓				✓
Google Images		✓		✓		
TinEye		✓		✓		
YouTube DataViewer		✓		✓		
FullFact	✓		✓	✓	✓	✓
LazyTrusth		✓			✓	
Holy Grail		✓		✓		✓
Trooclick	✓					✓

https://fullfact.org

Projets Check4Decison
Axes & Equipe de recherche

- Data Journalism
- Fact Checking
- Algorithme et Modélisation
- Traitement Automatique de la Langue
- Intelligence Artificielle
- Data Science

Equipe de Recherche et Partenaires

Prof. Ousmane SALL
Enseignant-Chercheur

Dr. Babiga BIRREGAH
Enseignant-Chercheur

Dr. Edouard Ngor SARR
Enseignant-Chercheur

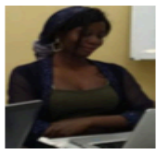
Prof. Mamadou Bousso
Enseignant-Chercheur

Dr. Marie NDIAYE DIOP
Enseignante-Chercheure

Dr Lamine FATY

Sény N. Mbaye (UT)

Mme. Senghor (UCAO)



Aminata MAIGA (UT)

Lamine FATY (UASZ)

Mory M.Koulibaly(AIMS)

Tony Tona Landau (AIMS)

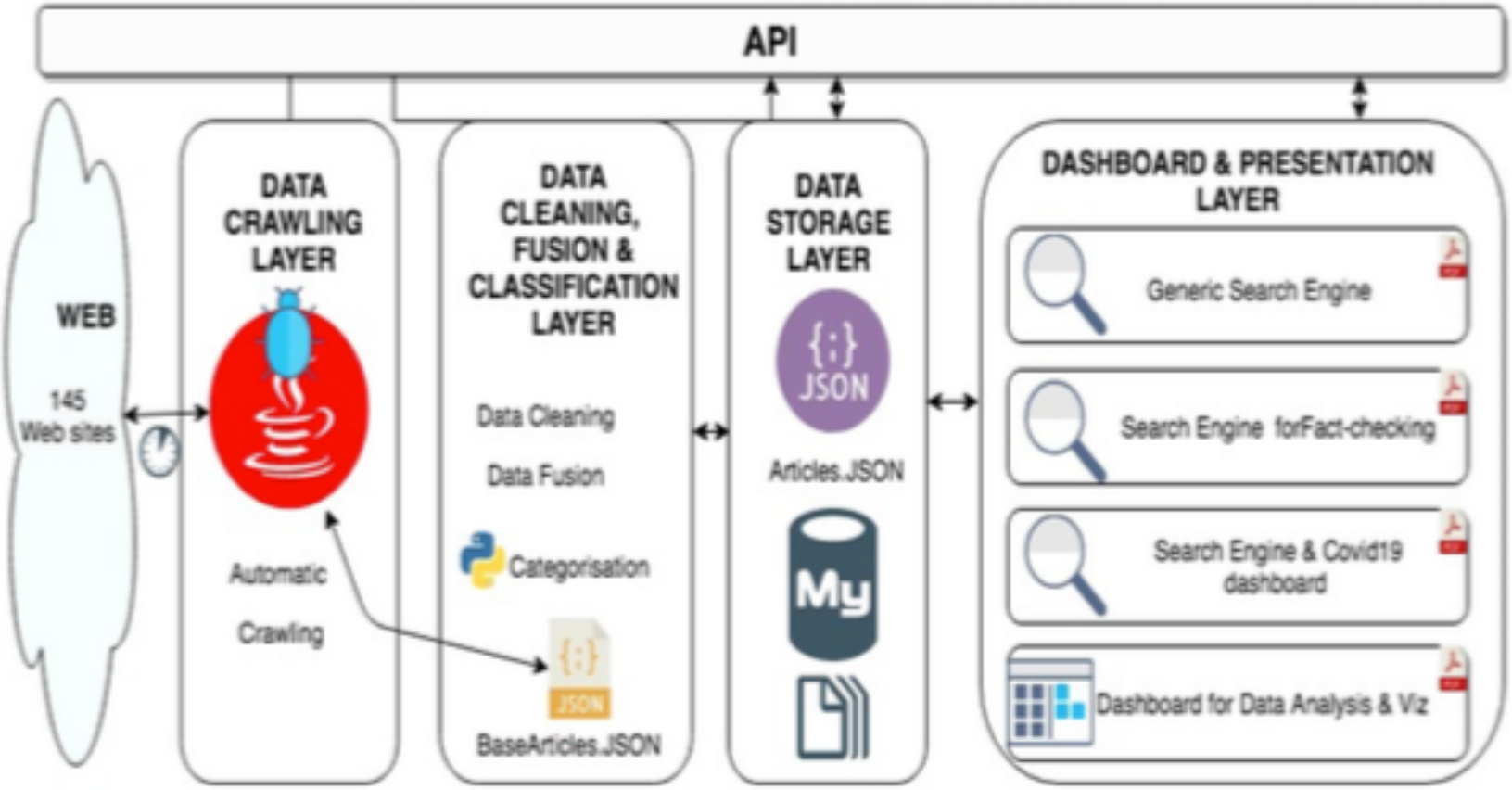
Moustapha Sissoko (UT)

Abdou L. Péne (UT)

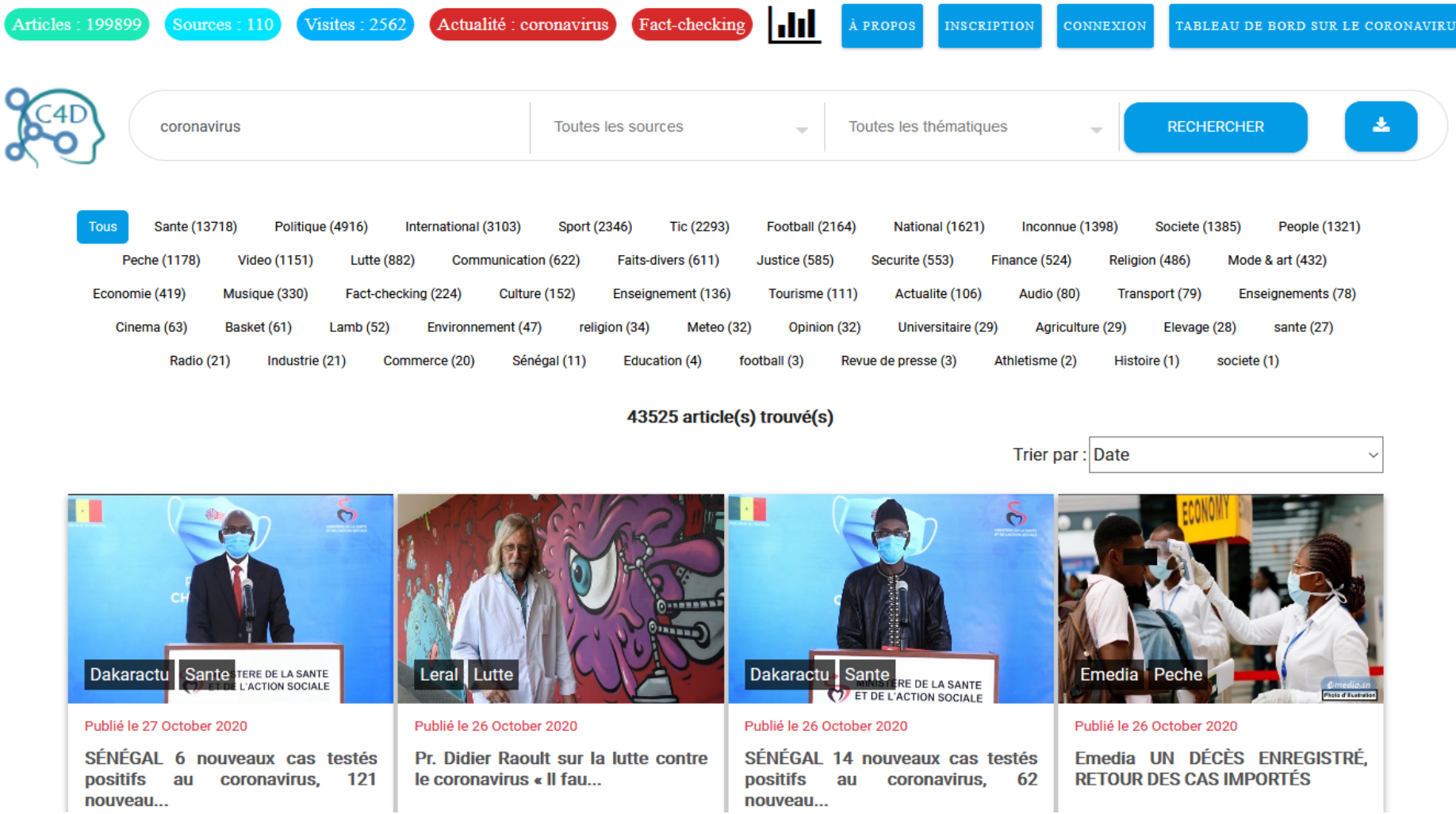
Fatima Toure

Projets Check4Decison
Architecture Générale du Système

Projet de recherche réunissant un consortium de 04 universités avec un financement initial du CEA MITIC http://check4decision.univ-thies.sn/

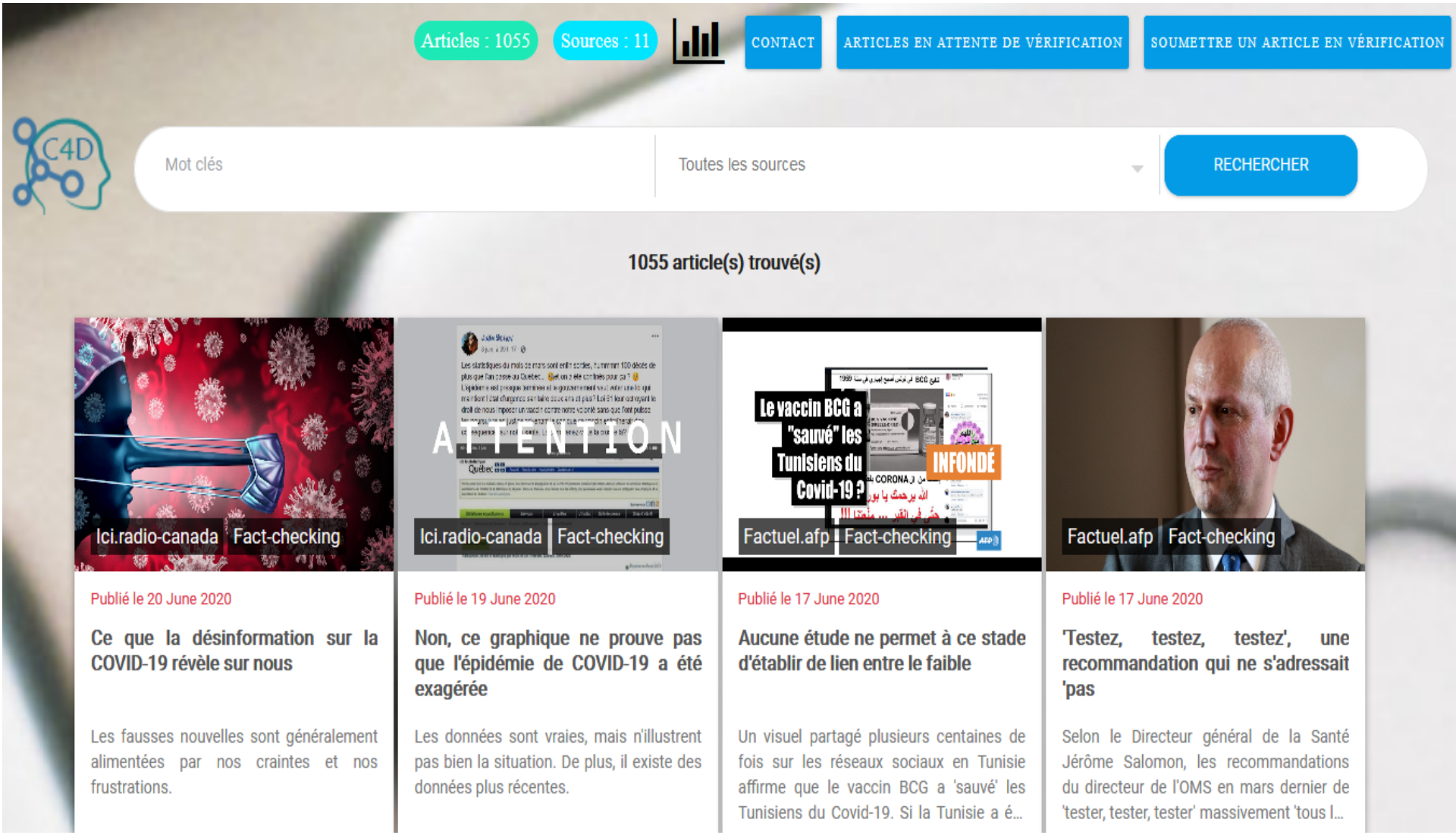


Projets Check4Decison
Résultats: Moteur de recherche



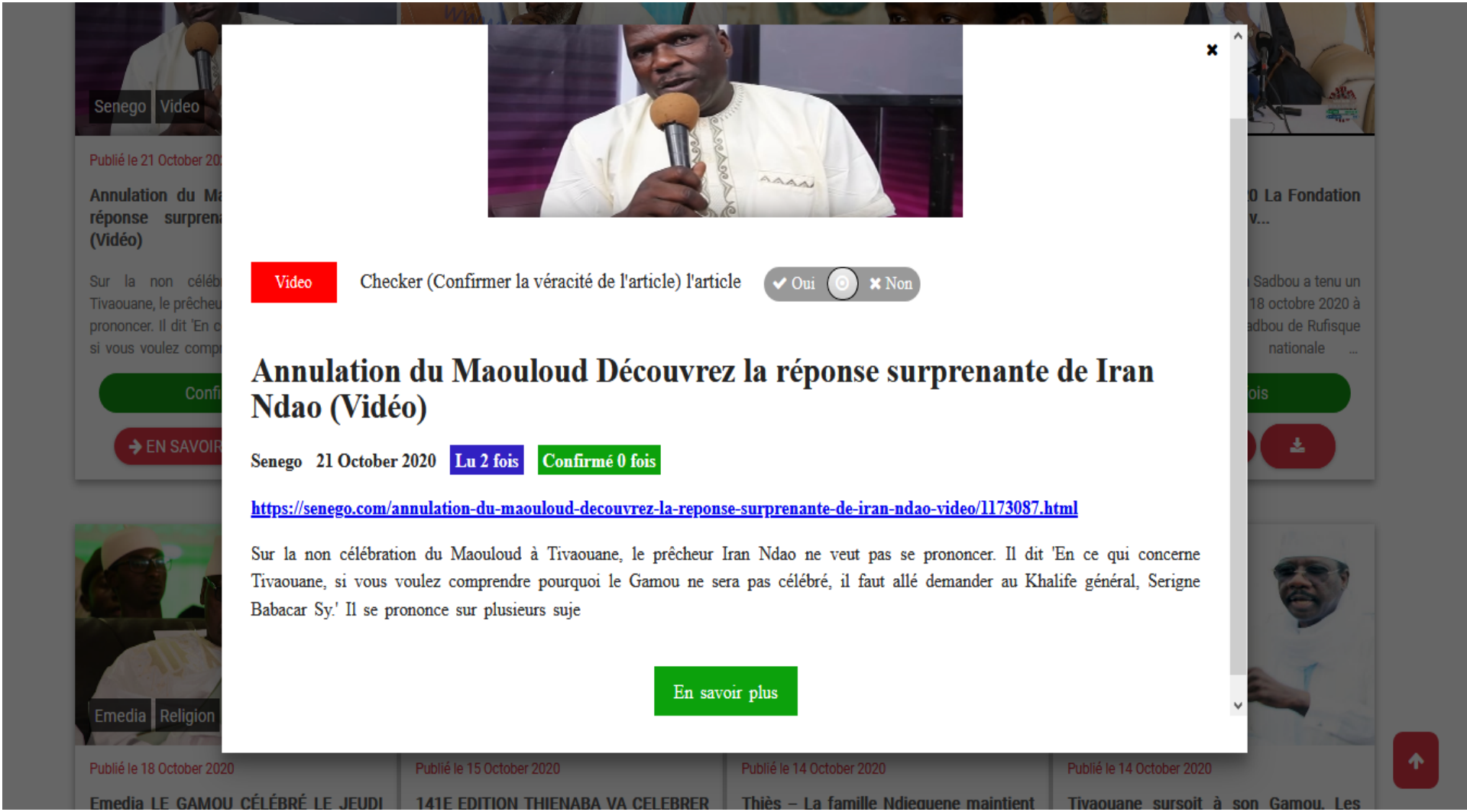
https://check4decision.univ-thies.sn/search

Projets Check4Decison
Résultats: Moteur Fact-Checking



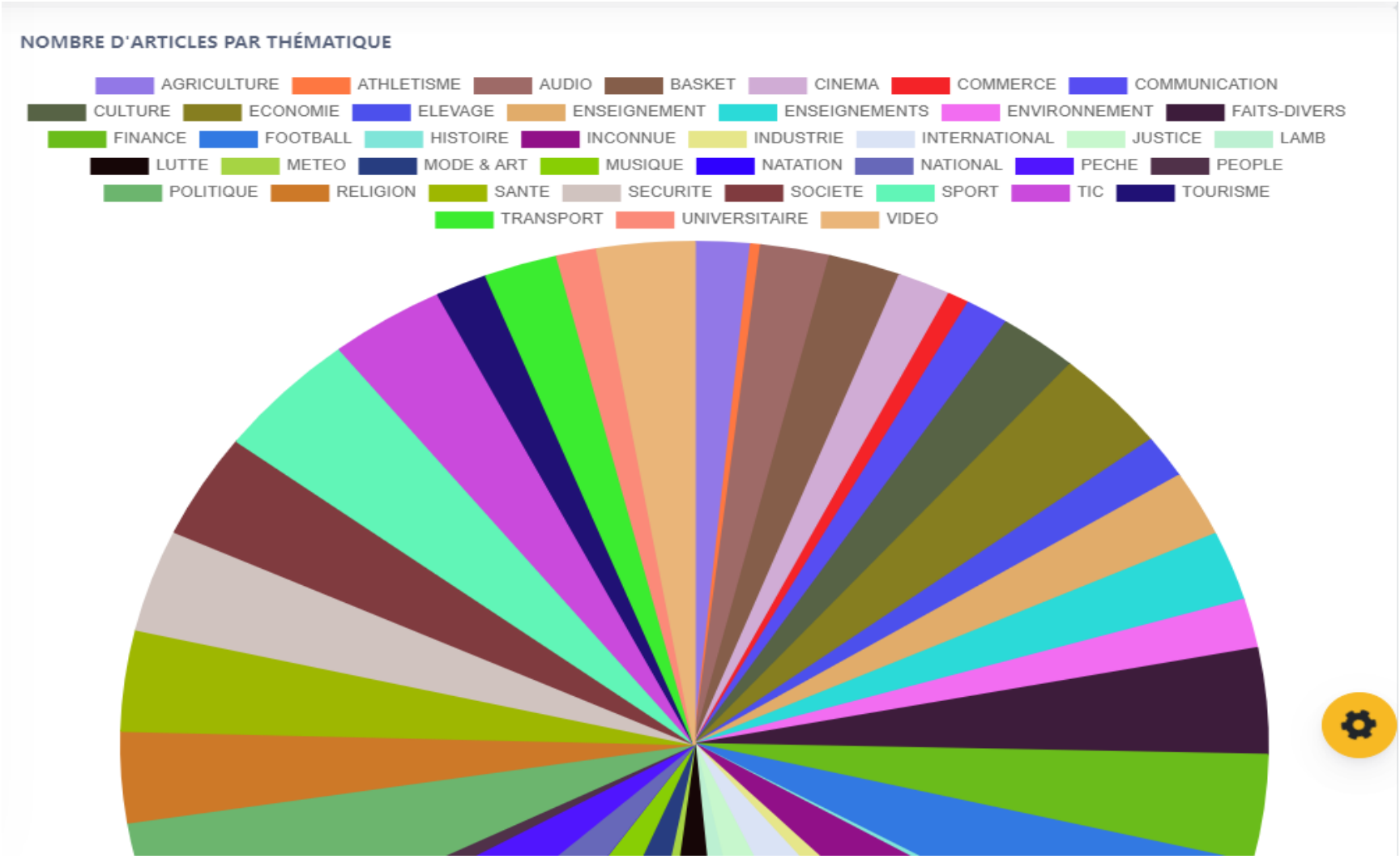
https://check4decision.univ-thies.sn/search/fact-checking.php

Projets Check4Decison
Résultats: Fact-checking Participatif



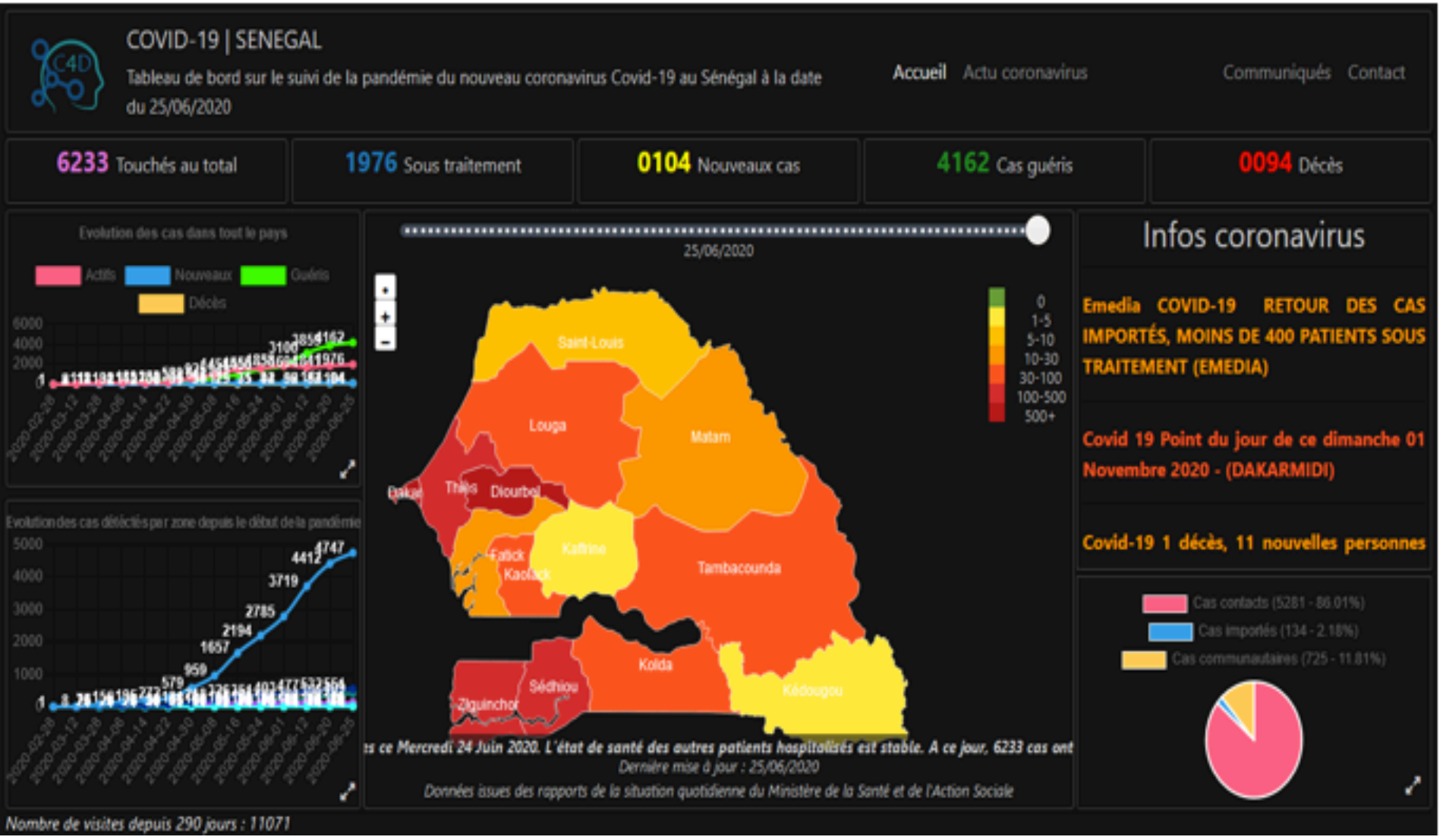
https://check4decision.univ-thies.sn/search/fact-checking.php

Projets Check4Decison
Résultats: Moteur Fact-Checking



https://check4decision.univ-thies.sn/search/login.php?page=statistiques-dashboard

Projets Check4Decison
Résultats: Moteur Fact-Checking



https://check4decision.univ-thies.sn/search/corona/

Projets Check4Decison
Travaux En cours

1. **Modélisation de la propagation des fakenews dans la presse en ligne**
   - Ed Ngor SARR & Dr Fulgence MANSAL, CERIDES-UCAO (Maitre Assistant, Spécialiste de la modélisation)
2. **Aspects légaux du web scraping et du crawling**
   - Ed Ngor SARR & Dr Philippe BASSENE, UASZ (Maitre de conférence en Droit, Spécialisé de la propriété intellectuelle)
3. **Identification automatique des faits dans les affirmations journalistique**
   - Thèse de doctorat de M Seyni MBAYE, ED2DS de l'Université de THIES
4. **Crawling Intelligent**
   - Mémoire de Master 2 de M Oumar DIAGNE, UASZ
5. **La classification automatique des articles de presse par apprentissage**
   - Mémoire de Master 2 de Mlle Fatima TOURE, FST UCAD
6. **Impacts des fakenews dans le choix des électeurs : le cas du Sénégal**
   - Mémoire de Master de M Landing (IMES-UCAO, Sciences Politique)

Check4Decision

Merci de votre attention

CEA MITIC